Standards-Driven Reform Years 1-10: Moderation an Optional Extra?

Val Klenowski
Queensland University of Technology

Claire Wyatt-Smith
Griffith University

Abstract

While externally moderated standards-based assessment has been practised in Queensland senior schooling for more than three decades, there has been no such practice in the middle years. With the introduction of standards at state and national levels in these years, teacher judgement as developed in moderation practices is now vital. This paper argues, that in this context of assessment reform, standards intended to inform teacher judgement and to build assessment capacity are necessary but not sufficient for maintaining teacher and public confidence in schooling. Teacher judgement is intrinsic to moderation, and to professional practice, and can no longer remain private. Moderation too is intrinsic to efforts by the profession to realise judgements that are defensible, dependable and open to scrutiny. Moderation can no longer be considered an optional extra and requires system-level support especially if, as intended, the standards are linked to system-wide efforts to improve student learning. In presenting this argument we draw on an Australian Research Council funded study with key industry partners (the Queensland Studies Authority and the National Council for Curriculum and Assessment of the Republic of Ireland). The data analysed included teacher interview data and additional teacher talk during moderation sessions. These were undertaken during the initial phase of policy development. The analysis identified those issues that emerge in moderation meetings that are designed to reach consistent, reliable judgements. Of interest are the different ways in which teachers talked through and interacted with one another to reach agreement about the quality of student work in the application of standards. There is evidence of differences in the way that teachers made compensations and trade-offs in their award of grades, dependent on the subject domain in which they teach. This article concludes with some empirically derived insights into moderation practices as *policy* and *social* events.

Introduction

This article reports on a four year Australian Research Council (ARC) Linkage[1] project, conducted in the Australian State of Queensland. The major industry partner for this project was the Queensland Studies Authority (QSA). Partnerships of academics researching in liaison with policy officers aim to inform policy development directly. It is often the case that policy officers have to "grasp the complex remit quickly and take action" (Saunders, 2005) without the benefits of policy-related research. The findings from this study were used developmentally to inform and influence state assessment policy related to the use of standards and moderation practices. This collaborative approach to policy reform investigated how policy was enacted during an initial phase prior to full implementation. The interactive nature of the research required investigation at the local professional level of districts and schools with ongoing feedback to the QSA at the political centre. QSA is the organisation responsible for the development of assessment policy at the state level.

Taken as a whole the ARC project addressed the following research questions:
How do stated standards work to inform and regulate judgement in different curriculum domains?

- What processes including social interactions do teachers rely on to inform their judgement decisions?
- What are the properties or characteristics of teacher judgements and how are these (as distinct from outcomes of grading decisions) shared or made available to other teachers?
- Does the social practice of moderation involving the application of explicitly-defined standards result in changed judgements about students' work?
- Does moderation using standards result in consistency of teacher judgement?

These questions are relevant to national and state education agendas across Australia where attempts are being made to develop a national curriculum (Reid, 2009) and national achievement standards in the subject areas of Mathematics, Science, History and English from K to Year 12. The research evidence from this ARC research has had bearing on the implementation of standards-based assessment locally and has implications for national and international assessment reform efforts. It is for this reason that our other industry partner was the National Council for Curriculum and Assessment in the Republic of Ireland.

More specifically this article explores teachers' judgement practice in the context of standards-driven reform with a focus on how the stated standards are used by teachers of different curriculum domains to inform and regulate their judgements. The processes and social interactions that teachers rely on to inform their decisions have been identified. The ways in which these teachers talked through and interacted with one another to reach agreement about the quality of student work in the application of standards have been analysed with evidence of differences in the way that they make compensations and trade-offs in their award of grades dependent on the subject area they teach.

In this article the issues that emerge in moderation meetings that are designed to reach consistent, reliable judgements are identified and empirically derived insights into moderation practices as policy and social events are reported.

The Changing Queensland Policy Context

For the first time in Queensland, teachers in the middle years of schooling (Years 4 to 9) are using defined standards to form judgements of student work and are engaging in social moderation. In the years of schooling from Pre-school to Year 10, teachers have in the past been required to use stated curriculum outcomes written as developmental markers, with a primary focus on their application to teaching and learning. They have not been required to use standards for assessing and grading purposes, nor has there been a requirement for them to undertake inter- or intra-school moderation as part of system efforts to support consistency of teacher judgement. The conceptual leap expected of teachers requires them now to assess student achievement on centrally-developed assessment tasks (Queensland Comparable Assessment Tasks; QCATs) using five defined standards (A to E), and to achieve consistency of judgement and reporting using the standards.

The Teaching and Learning Division of QSA devised and developed the Queensland Curriculum, Assessment and Reporting Framework (QCAR) that comprises the Essential Learnings, standards, an online bank of assessment resources, QCATs, and a Guide to Making Judgements that includes annotated samples of student work (see Table 1). Findings from this study have informed the development and refinement of some of these resources designed to assist teachers in using standards and moderation practice.

Threats to Validity in Developing Assessment Policy

At the time of this study QSA's P-12 draft assessment policy stated that there were six requirements for social moderation to work effectively. Two of the requirements pertinent to this research were that "syllabuses … clearly describe the standards of learning and standards of assessment" and that "teacher discussions of the quality of the assessment instruments and the standards of student work" (QSA, 2008b, p. 1) take

Essential Learnings	To identify what should be taught (key knowledge, facts, procedures and ways of working) and what is important for students to have opportunities to know, understand and be able to do.
Standards	To provide a common frame of reference and a shared language to describe student achievement.
Online Assessment Bank	To support everyday assessment practices of teacher through access to a range of quality assessment tools.
Queensland Comparable Assessment Tasks (QCATs)	To provide information on what students know, understand and can do, in a selection of Essential Learnings. QCATs are intended to promote consistency of teacher judgements across the state.
Guidelines for Reporting	To support consistency of reporting across the state.

Table 1: The QCAR Framework (QSA, 2008a)

place. This draft policy emerged in April 2008 from the Student Achievement Division of QSA concurrent with the QCAR initiative a responsibility of the Teaching and Learning Division. The timing and implementation of the assessment policy and the development of the QCAR framework (QSA, 2008a), which were the responsibilities of different divisions in the organisation, posed threats to the validity of the assessment process.

The QCATs were being designed, trialed and developed by the QCAR team using classroom teacher input and feedback. The tasks were written to provide an opportunity for “authentic assessment” as opposed to more paper and pencil test formats. Due to the pressures of timelines, budgets and product expectations the alignment of the teaching of the Essential Learnings (ELs) with the constructs of the QCATs was not always possible because the ELs were being introduced at the same time as the tasks were being trialed, developed and administered.

Identifying the key constructs for teaching and learning and subsequent assessment is fundamental to the concept of validity. All assessments are based on a sample of behaviour or performance in which we are interested and it is from the sample that we generalise to “the universe of that behaviour”. The “fidelity of the inference drawn from the responses to the assessment is what is called the validity of the assessment” (Nuttall, 1987, pp. 110-111). This is why the specification of the domain of behaviour in which we are interested is critically important. Our study revealed that some teachers had not taught the constructs of the QCATs yet they had administered the tasks to their students before they had the opportunity to learn these underlying constructs. This lack of alignment between the teaching of curriculum content and the use of tasks and standards raises issues related to validity. The analysis identified how this lack of alignment posed a threat to the validity of the assessments.

Crooks, Kane and Cohen (1996) suggest that developers of assessment tasks should address the threats to validity such as those linked to the scoring or grading of the student's performances on a task. These they identify as: the scoring or grading fails to capture important qualities of task performance; there is undue emphasis on some criteria, forms or styles of response; there is a lack of intra-rater and/or inter-rater consistency, and the scoring or grading is either too analytic or too holistic.

Further threats to validity include construct representation and construct variance

Construct representation refers to the extent to which the task samples the knowledge, skills and/or constructs that are being assessed. When "the test [or task] is too narrow and fails to include important dimensions or facets of the construct" there is construct under-representation (Messick, 1989, p. 34).

Construct irrelevance refers to the construction of the task and the reliability of the results (Messick, 1989). Construct-irrelevant variance exists when the "test contains excess reliable variance that is irrelevant to the interpreted construct" (Messick, 1989, p.34). This form of construct-irrelevant variance is regarded as a contaminant with respect to the score or grade interpretation. If teachers vary in their interpretation of the construct of interest that the task is designed to assess then there will be variation in the judgements and this will impact on the reliability of the grades awarded. Another example of construct irrelevant variance would be when the teacher's prior knowledge about a student influences the judgement made regarding the grade. To illustrate, if a teacher views a student in terms of their identity as an A student, and draws upon this knowledge when awarding the grade for the QCAT, then this is irrelevant to the construct that is being interpreted in the QCAT. That is, the focus should be on the work that the student has completed for the particular QCAT, not what has been accomplished previously.

These threats to validity imply that teachers need to be aware of the key constructs to facilitate their judgement practice. As indicated previously it was difficult for some teachers to be aware of the Essential Learnings which included the constructs as they were not implemented at the same time as the QCATs were being trialed and developed and at the time when this research was taking place.

Theoretical Underpinnings of the Study

The theoretical framework developed for this study incorporated a sociocultural understanding of learning involving Kress' (2000) view of language as inherently social and cultural and Sadler's (1987) work related to standards. A sociocultural view of learning sees learning as socially negotiated and embedded within a cultural community. We understand that learning is both a process of "becoming competent and belonging" to a community (Murphy and McCormick, 2008, p. x). Learning from this view is a "movement deeper into practice". It is also a "transformation of identity, where identity is understood as evolving forms of competence" (Murphy and Hall, 2008, p. ix). From this view "identity and knowledge are seen as interdependent". This sociocultural framing was adopted to analyse how teachers were becoming more competent in their consistent use of standards and how their identity was developing in the context of using standards to assess students' performance on the QCATs. The importance of context became apparent. The moderation meetings focused on the quality and standard of the student work being discussed. As the teachers participated in such meetings and contexts they drew on a range of concrete referents, social and tacit knowledge. At the same time they were learning in practice (about the standards) and negotiating an identity (as an assessor using standards).

By viewing language itself as inherently social and cultural (Kress, 2000) we understand that the terminology used to articulate the standards develops meaning over time, through use in contexts, including institutional and domain contexts and interactive contexts such as moderation practice. The latter can take place incidentally or more formally in a structured manner through system and school initiatives. The teacher or sign-maker is "constantly transformative of the set of resources of the group and of her/himself" (Kress, 2000, p. 401). "You just have to learn how to see" applies to teachers when they use standards to assess student work to achieve consistency in judgement during moderation practice. Our analysis revealed multimodal interaction with several modes of representation and communication when exchanges took place between teachers in the moderation meetings. In the different subject domain meetings there appeared to be a specialisation of representational and communicational modes during those interactions.

According to Sadler (1987) standards-referenced assessment relies on teacher judgement that can be made dependable if standards are promulgated in appropriate forms and teachers have the conceptual tools and professional training. The methods for promulgation include: numerical cut-offs, tacit knowledge, exemplars and verbal descriptors. Sadler argues that when multiple criteria are used in assessment (as is the case in the QCATs) it is more viable to work with "fuzzy composite standards" (Sadler, 1987) in contrast to sharp standards characterised by precise boundaries and numerical cut-offs. "[V]erbal descriptions are always to some degree vague or fuzzy" (Sadler, 1987,

p. 202). The fuzziness of grade boundaries comes from the interpretation on a particular criterion underlying the standards scale *and* the interpretation of the different levels separating one standard from another. Both these interpretations depend on context. A competent assessor using verbal descriptors to judge student work will make "compensations and trade-offs … to allow for intercorrelations among the criteria and for the multiplier effect of some criteria on others" (Sadler, 1987, p. 206). The teachers in this study were required to use standards that were written as verbal descriptors and therefore they had to interpret from the evidence in the student work which standard to award. Such interpretations we found were indeed influenced by context and the subject domain.

Sadler (1987) makes reference to Wittgenstein (1967/1974) who understood that precise definitions for a lot of the everyday concepts was impossible. In his view they comprised "a complicated network of similarities, overlapping and crisscrossing: sometimes overall similarities, sometimes similarities of detail … 'family resemblances'". A combination of language and experiences with real phenomena help "sketch in" meanings. Sadler (1987, p. 206, emphasis added) concludes, "the verbal description of a standard (the standard itself being an abstract mental construct) can have its interpretation, circumscribed, more or less adequately, *only by usage in context.* The concrete existential referents that make up the context are essential to its proper interpretation".

The study has identified an empirically derived framework of categories of referents that teachers have used to confirm or challenge proposed grades during the moderation meetings (Wyatt-Smith, Klenowski, & Gunn, 2010). These two categories of referents include, first, the textual materials provided by QSA (i.e., the products such as the Guide to Making Judgements, or the Guide, and the annotated student work samples), together with the readings or interpretations that the teachers have made of these. The second category includes the types of tacit and social knowledge evident in the recorded talk but missing from the official documentation. Data analysis revealed that these include discipline (subject) knowledge; knowledge of the official curriculum; the teachers' prior evaluative experience; knowledge of individual students, and knowledge of what the average student could reasonably be expected to demonstrate at a given year level.

Methodology

A qualitative methodology was adopted and the data collection methods included pre- and post-moderation interviews of focus teachers, observation and collection of conversation data from an initial four cluster moderation meetings. These first meetings involved a total of 15 schools, 11 state schools and four independent schools. Fifteen focus teachers were interviewed before and after the moderation meetings. Their talk was also tracked during the recorded moderation sessions. The following table provides a brief overview of the composition of this group of focus teachers.

Gender	School Level/Subject Area	Sector
12 Female	Primary 8 Year 4 (English) 2 Year 6 (Maths)	11 State
3 Male (all secondary)	Secondary 4 Year 9 (Science) 1 Year 9 (Maths)	4 Independent

Table 2: Composition of the Group of Focus Teachers

This data was supplemented in 2007 by a further 11 schools that trialed the QCATs. Data collection for this year of the study included a further 35 focus teachers with approximately 150 teachers involved in moderation talk session across the three Key Learning Areas (KLAs) of Mathematics, Science, and English in Years 4, 6, and 9. All interview and meeting data were progressively transcribed in full for analysis by more than one researcher.

Discussion and Findings

What follows is an analysis of data that illustrates how teachers are learning to see, or becoming competent in their use of standards, from their interpretation of the grades and their use of textual materials. During the interactions of the moderation meetings there are expectations expressed, as they were in the pre-moderation interviews, of verification, vindication and validation of judgements. There is also evidence of the teachers being transformative of the set of resources of the moderation meeting and him/herself. The modes of representation and communication are speech used as an oral mode of commentary, critique and ratification, as well as written comments recorded on student samples, annotations entered on the Guide, usually involving "shorthand markers" of a quality such as ticks or highlighting, and in some cases a written comment to accompany the letter grade.

Standards informing judgement

Standards have been categorised into five types (Maxwell, 2008) but for the purposes of this paper the two that are most relevant are those concerned with differentiated levels of performance. Standards described as "arbiters of quality", indicate relative success or merit, the other type is described as milestones and are indicative of progressive or developmental targets. These two types differ in focus and timeframe. The former are merit standards that can apply to a single assessment event, such as a completed task or the QCAT, while the latter are developmental standards that inform judgements that can be made along a continuum of learning over time, such as for a portfolio of work (Maxwell, 2008).

In the following pre-moderation interview Chris indicates what he hopes to gain from the moderation meeting. He describes how he might "learn to see" the standard as he has interpreted it. This is a merit standard or an arbiter of quality so the focus in the judgement is on relative success or merit of the student's response. The importance of context, the interactions and the transformative nature of the available resources become apparent:

> Chris: I think I'll be interested to see, … the most important part for me … will be meeting the teachers and hearing what they say, and seeing also, for my own thing, seeing if, I mean, I'll go along there with a pre-conceived notion already of what those, the QCATs and thinking, "Well, okay, I think so-and-so's going to be an A," so in the back of my mind I'll have their work. I'm really interested to see whether that matches up with what other teachers think about it, too. And if it does it vindicates me. If it doesn't I have to go back and say, "Okay, I am marking too hard, I am marking too easily." So that will be good, that will be really good.

The importance of the interaction during the moderation meeting appears to be acknowledged by this teacher as he reveals how his grading of the student's work and his expectations will either be confirmed or challenged. The interactions are important in the transformation that is anticipated by this teacher through the resources of the group and the teacher's subjectivity. His interest in aligning his judgement with that of other teachers is apparent when he states that in matching up his judgement with those of other teachers the outcome will "vindicate" him.

In what follows a teacher identifies the criteria (assessable elements) related to the science QCAT, and the task itself, as problematic in informing judgement due to the ambiguity associated with the construct to be assessed. The teachers agreed through dialogue to communicate first their dissatisfaction with the task or "instrument" and second the transformative aspect of the referents available to the group and its members. In line with Kress' (2000, p. 401) "theory of the constant transformation of both resources and of subjectivity" here we see individual group members being agentive in relation to the group's resources, and in relation to the individual's own subjectivity. Don explains how teachers are forced to look seriously at the resources that are available to the individual in that transformative activity.

Don: I think it was very much centring on the way in which the assessable elements are written... So basically, I think, all of our disagreements, to put it in its essence, were the different ways in which teachers tried to adjust for perceived weaknesses in the instrument and its criteria.

I: Right.

Don: Now, the last question talks about floating, sinking, Plimsoll line, weight and force. And then the assessable element out of nowhere grabs upward force. Upward force is not in the stimulus material at all. But to get an A or a B you had to use upward force, so we just ignored it. And we agreed to ignore it. But we had to dialogue the fact that we were ignoring part of the instrument because it's strop[2].

These data are illustrative of the key threat to validity, of construct under-representation, as, according to the teachers, the task does not appear to adequately sample the construct of upward force that it is supposedly designed to assess. The important finding here is the need for greater clarity and explication of construct definition in the task for teachers.

Use of standards for improved learning

The standards were seen to be beneficial in that they helped some teachers and their students to focus on the qualities that were assessed in the completed work and they helped to address the threats to validity such as construct irrelevance or construct under-representation. All teachers did not share this view. In what follows, Carl suggests that the standards help to crystallise the qualities of the work for both the teacher and the students to see how the standards focus student attention on "what they need to improve and what are their strengths" and "what we have marked". However, the use of standards for learning does not appear to be part of Carl's pedagogy suggesting that he has used the discourse without fully understanding the implications for practice. This is an example of "false clarity" in that Carl appears to see the connection between criteria and standards and learning improvement but does not seem to know how to realise this pedagogically.

I: … what are the benefits of using stated standards in your opinion?

Carl: Um, it gives a much more lucid and crystallised idea of … what the qualities of this work are and what, … the deficiencies might be as well, and … it's much more based on the work rather than the student. … it's also a transparent system … it creates much more equity between students because it is entirely task-based and it is the quality of the work that we are assessing at all stages … Detail is important as well. Without that detail it becomes murky and open to interpretation … it gives the students a much better idea of what they need to improve and what are their strengths, um, and they can use … those statements of standards and what we have marked and how we have used that page in order to have a much better understanding of what, further than our comments or just having a mark, of what they have achieved.

Carl does not appear to know how to assist students in the use of standards for learning by incorporating assessment for learning strategies. His talk about criteria does not include the provision of opportunities for students to apply the criteria. For example, his pedagogy does not extend to teacher modelling of how to use the criteria for self-monitoring and improvement purposes and exemplification through illustrative samples of student work.

I: Okay. Now, so then what are your concerns about the use of stated standards?

Carl: … one of my major concerns is whether or not the students can actually use them, whether they are a tool for students or whether they are a tool for, for teachers. And even with many of my attempts with my classes, for example, the students do not find much currency in them, despite the fact that we have pored through it and we have said, "This is what you need to achieve and this is exactly what an A would be. … this is where you would need to, this is what would allow you to achieve that standard" … And that still doesn't hold very much currency for the students...

Tony in contrast to Carl can see more benefits for students in their use of the standards for self-assessment to identify the gains that they have made or the areas where they need to develop or focus.

Tony: Um, I can see benefits for the kids in that they can actually see, um, what requirements they need to obtain a certain level, or an A, B, C or D, so basically they can go through their work and say, "Yep, now I've done that, I've done that, I've done that, I've done that. I should be getting around about this mark when I get it back." Or they think, "Oh, geez, I should have, I could have, I haven't really done that so that's going to bring me down."

The need for consistency of the application of standards for student improvement is a recurring theme across the data sets and illustrates the acceptance by teachers of the underlying principle inherent in the policy intent of the QCAR initiative. Consider for example the clear resonances in what Carl, Tony, Ian and Cathy say regarding stated criteria and standards across these segments.

Ian: Um, … the point of the standards is to help, to help achieve consistency, and I think consistency is important because it helps students know what they need to do. … the worst thing is when you have an assessment item and you don't actually know what's expected of you. You know, "Is this enough? Is this too much? How do I know?"

Cathy: I think it's very clear to both staff and student at the end of a task, and therefore at the end of a year and looking at all of those tasks, they know what standard they're at. … I've always been a big believer in standards. I just think it gives a lot more security to the students in their learning and to the staff in the delivery of the teaching...

From the analysis it appears that, in the main, the teachers' accounts of policy in practice align with the official policy directions and partial uptake of the QCAR initiative; however, there is still the need to build capacity in teachers' understanding of assessment as it connects with learning theory.

Teacher Judgements: Analytic and Holistic Approaches

The role of criteria and standards

The textual representation of assessment criteria and standards or the Guide directs teachers in particular ways to understand markers of quality. Such understanding is expected to operate at the micro level where the focus is on discrete assessable elements (criteria) linked to questions of the QCAT. It also operates at the macro level that involves the award of an overall grade. In judging student achievement on the centrally-devised QCATs, teachers were asked to arrive at "on balance" judgement. This required attention to qualitative levels of difference, both across the completed tasks to be assessed, as well as among dimensions of performance within a task.

In the QCAR initiative, priority was given initially to the micro level. The Guide for Year Six Science, in its original design adopted a matrix approach, and the annotated student work samples provided meticulous specification of what teachers were to do to assess the student's work. To illustrate:

> locate the evidence in the student work for each assessable element. Match the evidence for each assessable element to a task-specific descriptor in the Guide to Making Judgements. Refer to the Annotated student work samples (if available) to support your understanding of the expected student response for each task-specific descriptor. (QSA, 2007)

Teachers were thus informed to assess by judging the component parts of the work against each element of the Guide, represented in matrix format, using annotated samples as support. Teacher judgement was in this way oriented to an analytic approach, focusing on prescribed, discrete elements. The assumption was that the process of treating each element separately, and in turn, would lead to a systematic, even regulated approach to judgement that could deliver consistency. Brief notes on obtaining an overall grade were available, however no exemplars were available to illustrate the qualities expected for overall final grades (A – E). The criteria were atomised at the level of the question and the standards (task-specific descriptors) and were intended to assist in the overall judgement for the award of a final grade however they remained in the background.

Discipline differences and their role in shaping judgement

The talk recorded in interviews and moderation meetings also brought to light the contributions of discipline knowledge in teachers' expectations about how judgement should properly occur. Specifically, the data showed how teachers of English, for example, relied on judgement processes that were significantly different from those of mathematics and science teachers. This was the case even when they were using a Guide to making judgement that had a common design across the three discipline areas. This points to how teachers' ways of recognising achievement can be traced directly to particular constructs of knowledge.

Consider, for example, how in answer to the question about the procedures and processes relied on to achieve consistency of teacher judgement in science, a teacher refers to the benefits of "following a similar marking key… a very detailed marking key". For teachers of mathematics and science, the recurring interest was in how tightly specified marking information was essential in regulating judgement, and moreover, that questions should be accompanied by marks that indicated their relative importance in the assessment item. In the following interaction, a teacher of primary maths and science was asked to elaborate on his statement that "the key thing [for achieving consistency of teacher judgement] was to get a common understanding of the standard", to which he said:

> In the, in the ideal world of education we have the task, we have its criteria and we have a sample response. Ah, in maths and science exams that's easy, you have a marking sample. And the other teachers can look and go, "Okay, so there's two marks for that question and that's broken down by identifying the formula." So it's not just seeing the marks. I would also expect teachers … to also be able to see the part marks.

In this segment the teacher can be heard associating criteria, a sample response and a marking schedule that stated how marks, including part marks, were to be awarded. The concern is with the component parts of the work to be assessed, though missing from the talk is how marks and standards connect in practice or conceptually.

Overall, the prevailing assumption in the recorded talk of maths and science teachers was that marks were used not only to regulate how judgement should occur, but that marks were the building blocks for arriving at overall judgement. In short, the teachers' assessment gaze was with the parts and that the whole necessarily represented the sum of the parts. Moreover, there was the expectation that judgement practice could be wholly pre-specified, some teachers talking about the merit of prescribing the answers to be recognised for half and quarter marks. One teacher spoke of this saying how teachers "really have to be more consistent in, where some teachers might be inclined to give a half-mark, um, where others might only give a quarter mark – that kind of thing. I usually find that that's the area that needs the most attention".

Further, from maths and science teachers there was a level of reported discomfort in using standards that they regarded to be "open to interpretation". In the words of one teacher, the matrix approach to specifying standards, referred to above, " is obviously open to interpretation". Continuing on, the teacher characterised the matrix as "probably not as objective as it needs to be", commenting favourably on how in the school, "we make up a proforma – a matrix for marking and marking criteria". It was as though standards written as verbal descriptors, representing fuzzy standards, needed to be fixed and that a marking criteria using numeric scores needed to be developed to stabilise the meaning of the terms used in the standards.

Such practices, while common in science and mathematics, were not evident in English. One teacher captured the widely reported judgement stance in English as follows:

> We're always looking for a global assessment. We're not looking at a precise number in English, so you're always looking at a global…

Overall, English teachers tended to move from the whole to the part, preferring to regard the work to be assessed in its entirety, before fixing on particular aspects of performance. Further, collectively they voiced concern about "a danger of being too detailed" in specifying criteria and standards, as indicated in the segment below:

> They need to be, well, for English teachers who interpret every single letter, let alone every single word, they need to be explicit and they need to be up-front and they need to be well-understood by people prior, perhaps, to doing anything with them. Um, but you never really understand what they're about until you are grading or you are using them. So, until you see them in operation it's hard to know, but there is a danger of being too detailed and almost verbose with what you're trying to do ... The standards have to reflect really, it hones in therefore on what it is you're really assessing.

In this segment we hear the teacher disclosing how knowledge of set standards is developed through practice, and warning against an overly detailed approach to how they are formulated. There is also the telling statement that the standards "have to reflect" what is *really* being assessed.

Such an observation opens up for consideration how, broadly speaking, maths and science teachers, and English teachers, have different expectations of how standards function in informing judgement. For the former group, the standards in conjunction with numeric scores are expected to regulate teacher attention while for English teachers the given standards become meaningful through use. It is as though teachers try on the fit of the standards for student work as an evaluative experience, with the terms in which the standards are written developing meaning within a marking occasion, and from one occasion to the next.

The final observation in this section of the paper concerns the longstanding distinction between objectivity and subjectivity and the prevalence of these terms in teacher talk about disciplines and judgement practices. This study provided some considerable evidence that maths and science teachers are more likely to claim that objectivity of judgement is realised through numeric scoring. English teachers reported valuing holistic judgement, taking this to be subjective judgement practice. This is clearly voiced in the segment below:

> T: English is inevitably going to have some subjectivity in it, um, and then I think this is the challenge with defining our criteria in each of the standards… as I said, we want to be, we want to be as objective as possible, and um, that's very difficult in English. It's just not a you know…
>
> I: It's not an objective subject.
>
> T: That's right. You know, it's not a quantitative, you know, assessment. It's qualitative and quantitative, you know. It's the holistic judgement of the piece and, you know, you've got to find a balance.

The reference to English teachers wanting to be as objective as possible can be heard as signalling a commitment to fairness in grading. It can also be heard as signalling that while English teachers aspire to objectivity, in the opinion of this teacher, this is not realised through adopting what she refers to as a wholly regulated, quantitative approach.

The discussion to this point suggests that the ecology of teacher judgement is shaped as much by assumptions about the nature of knowledge as it is by any given or prescribed approaches to judgement using a Guide or set criteria and standards. This study also showed how judgement practice can be shaped by textual materials provided to teachers, as discussed next.

Teacher use of textual referents

As suggested above, the recorded talk showed that teachers relied on a range of practices and referents to make the move from the parts (micro) to the whole (macro) in arriving at a judgement. These included giving priority to the annotated samples, referring only to the Guide as a secondary source of information; giving each of the criteria a numeric sub-score and then totalling the sub-scores to arrive at an overall grade; parcelling out the marking to different teachers to judge certain sections of the paper only and then passing the responsibility for overall judgement to another party (usually a senior teacher or curriculum leader) to combine the judgements on the separate criteria into a composite grade. Martin can be heard indicating his reliance on the samples as a way to cue into categories primarily to inform judgement. Rhonda is also reliant on samples to inform her judgements.

I: So what process did the teachers go through in terms of reaching their judgements?

Martin: In terms of reaching their judgements the, um, the teachers had pretty carefully read through all of the information provided by, um, QCAR and they had made sure that they, um, firstly looked at the, they had a look at the tests themselves, they looked at the sample responses and they tried to, ah, align their judgements with the sample responses.

I: And then sort of mark those on the back with all of the, with what they call the Guide on the back?

Martin: Yes, which we found to be both helpful and unhelpful. In some cases, see, answers were, ah, responses were very explicit and they fitted into a category very easily. In other cases, students answered in different ways which didn't make categorisation easy.

Rhonda also explains how, in assessing the mathematics task, problems emerged for her when combining components of the task to award an overall grade. It was in this context that the teacher referred almost entirely to the sample responses to see how the grades were awarded. However, she found these samples did not account for all possible variations so that: "It was difficult to give a C when one question was fully right and one was fully wrong".

I: Yeah. So, how did you use the materials and, and what was the way you went through the task?

Rhonda: I basically used the task-specific descriptors[3] and the question numbers relevant to it and then graded it, as I told you before, I have done a lot of senior Maths marking where we mark on criteria, so

> that helped me a lot to do it, so I feel that was quite straight-forward and easy that way. Though, when marking individual questions, it was a bit difficult to give them a C when one question was fully right and one question was fully wrong and those types of things and we are to give how much value? That was a problem there.

Both Martin and Rhonda touch on the complexity of judging, suggesting their interest in materials that could make categorisation easy. There is some suggestion that the matrix approach of the Guide hindered the award of an overall grade because of the extent to which the QCAT and the assessable elements atomised the teachers' approach to judgement. A more experienced Head of Curriculum (HoC) spoke of how he drew on his evaluative experience – another way to see – in working with teachers in his department to show how strengths and limitations could be evidenced in a piece of work and how these could facilitate on balance judgements.

> Ben: I think what, ah, when we had a look at this QCAR products [sic], when we had a look at those and we saw the student responses to that, it was interesting in that with the A description, the B, I think that was a very good method of showing what is required. … But I had a chat with one of my other colleagues before and he was saying, "You know, this, this question here, that's an A standard," and then he turned the page and showed a B standard and the wording was virtually the same. I had to point out to him that what the standard was showing you wasn't just for that question but for the whole paper. So, you know, the B standard, to my way of thinking, for both questions, the two answers for both questions were a very good answer, but it was in the other questions where the B standard would have come out, not necessarily just in that one question. I think that's something that needs to be pointed out…

This HoC has demonstrated his awareness that in making an on balance judgement the approach requires "best fit" rather than "perfect match" in terms of the teacher's interpretation of the evidence demonstrating the student's understanding of the construct being assessed at a particular standard, in this case a "B".

Research Implications

The findings of this aspect of the overall ARC study indicated that a common interpretation of the standards, at the level of chosen discipline tasks, was in development. Also clear was that, in the main, teachers had not connected the Guide to the standards developed as part of the QCAR initiative for judging achievement at the discipline level. That is, there was no conceptual bridge linking the Guide to discipline standards. Further, it is worth emphasising that while it is widely recognised that discussion among teachers regarding the evidence depicting the qualities of standards is fundamental, our observation has been that such discussion will not necessarily occur in the absence of policy direction. We have suggested that this observation holds, even if individual schools are proactive and make time provision for moderation linked to professional learning.

These research findings informed the next stage of the QCAR implementation in that an alternative design for the Guide was trialed. The alignment of standards to criteria was shown graphically on continua as in the work of New Basics (Klenowski, 2007). This formed part of our continuing research.

This study also helped us to report on how to develop teacher assessment capacity in the use of criteria and standards to inform judgement for teaching, learning and reporting purposes. A professional development strategy that includes assessment and moderation principles, judgement approaches and accompanying resources has been developed. Teacher use of standards, at both task and discipline levels for application at National and State curriculum and assessment priorities, has been the focus.

Acknowledgements

The authors wish to acknowledge the funding received from the Australian Research Council and the involvement and support provided by the Queensland Studies Authority (QSA) and the National Council for Curriculum and Assessment of the Republic of Ireland as our Industry Partners in this research project. Other members of the research team include: Prof J. Elwood, Dr A. Looney, L. Adie, S. Gunn and P. Colbert.

Endnotes

[1] The project is funded by the Australian Research Council in collaboration with Industry Partners, the Queensland Studies Authority and National Council for Curriculum and Assessment (The Republic of Ireland).

[2] In this context "strop" means "frustrating" because the construct of upward force is lacking in the stimulus material.

[3] The teacher has used the incorrect term in this context. The assessable elements identified the constructs and their relationship to the questions of the QCAT, not the task-specific descriptors which were in fact the standards. That is the task-specific descriptors are the standards and the assessable elements are the constructs and/or criteria.

References

Crooks, T. J., Kane, M. T., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, *3*(5), 265-285.

Klenowski, V. (2007). *Evaluation of the effectiveness of the consensus-based standards validation process*. Townsville: DETA. [Available online: http://education.qld.gov.au /corporate/newbasics/html/lce_eval.html]

Kress, G. (2000). "You've just got to learn how to see": Curriculum subjects, young people and schooled engagement with the world. *Linguistics and Education*, *11*, (4), 401-415.

Maxwell, G. S. (2008, September 7-12). *Setting standards: Fitting form to function*. Paper presented at the 34th IAEA Annual Conference, Cambridge UK.

Messick, S. (1989). Validity. In R. Linn (Ed.) *Educational Measurement* (3rd ed.), pp. 13-103. New York, NY: American Council on Education and Macmillan.

Murphy, P., & McCormick, R. (2008). *Knowledge and practice: Representations and identities*. London: Sage.

Murphy, P., & Hall, K. (2008). *Learning and practice: Agency and identities*. London: Sage.

Nuttall, D. (1987). The validity of assessments. *European Journal of Psychology of Education*, *11*, 109-18.

Queensland Studies Authority (2007). *Reviewing Process*. Brisbane: Queensland Studies Authority.

Queensland Studies Authority (2008a). *The QCAR Framework*. Retrieved March 25, 2008, from QSA website, http://www.qsa.qld.edu.au/assessment/qcar.html.

Queensland Studies Authority (2008b). *P-12 Assessment Policy*. Brisbane: Queensland Studies Authority.

Reid, A. (2009). Is this a revolution? A critical analysis of the Rudd government's national education agenda. *Curriculum Perspectives,* 29(3),1-13.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191-209.

Saunders, L. (2005). *Policy Research*. Presentation at the Institute of Education, University of London.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. J. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education*, *17*(1), 59-75.